

# Semi-automatic selection of summary statistics for ABC model choice

Dennis Prangle<sup>\*†</sup>, Paul Fearnhead<sup>‡</sup>, Murray P Cox<sup>‡§</sup>, Patrick J Biggs<sup>‡¶</sup>  
and Nigel P French<sup>‡¶</sup>

February 25, 2013

## Abstract

A central statistical goal is to choose between alternative explanatory models of data. In many modern applications, such as population genetics, it is not possible to apply standard methods based on evaluating the likelihood functions of the models, as these are numerically intractable. Approximate Bayesian computation (ABC) is a commonly used alternative for such situations. ABC simulates data  $x$  for many parameter values under each model, which is compared to the observed data  $x_{\text{obs}}$ . More weight is placed on models under which  $S(x)$  is close to  $S(x_{\text{obs}})$ , where  $S$  maps data to a vector of summary statistics. Previous work has shown the choice of  $S$  is crucial to the efficiency and accuracy of ABC. This paper provides a method to select good summary statistics for model choice. It uses a preliminary step, simulating many  $x$  values from all models and fitting regressions to this with the model as response. The resulting model weight estimators are used as  $S$  in an ABC analysis. Theoretical results are given to justify this as approximating low dimensional sufficient statistics. A substantive application is presented: choosing between competing coalescent models of demographic growth for *Campylobacter jejuni* in New Zealand using multi-locus sequence typing data.

*Keywords:* ABC, model selection, sufficiency, *Campylobacter*, MLST, coalescent

---

<sup>\*</sup>d.prangle@lancaster.ac.uk

<sup>†</sup>Department of Mathematics and Statistics, Lancaster University, UK

<sup>‡</sup>Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

<sup>§</sup>Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

<sup>¶</sup>Infectious Disease Research Centre, Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand

# 1 Introduction

The increasing availability of modern genetic data offers the possibility of learning more than ever before about the processes which generated it, for example the details of demographic change. However, for stochastic models that incorporate a high level of detail, it is impractically costly to evaluate numerically the probability of a dataset, preventing inference by standard likelihood-based methods. This has motivated the development of likelihood-free approaches, such as approximate Bayesian computation (ABC), which utilise the fact that simulating data from these models is relatively computationally cheap.

There is particular interest in using these methods to choose between explanatory models for observed data. However Robert et al. (2011) illustrated that applying ABC to model choice problems can produce highly inaccurate results. This paper provides methods to address these concerns and improve the informativeness and efficiency of ABC model choice. We focus on a particular application, inferring the demographic history of *Campylobacter jejuni* in New Zealand from population genetic data. This will be described in detail later.

A simple ABC algorithm operates by simulating data sets  $x$  under various model and parameter pairs  $(\mathcal{M}, \theta)$ . Pairs are accepted when  $x$  is sufficiently close to the observed data  $x_{\text{obs}}$ . This produces a sample of independent draws from an approximation to the Bayesian posterior distribution i.e. that of  $\mathcal{M}, \theta | x$ . Closeness is judged by the distance between vectors of *summary statistics*  $S(x_{\text{obs}})$  and  $S(x)$ . Previous work (e.g. Blum 2010; Fearnhead and Prangle 2012) has shown that the quality of the approximations produced by ABC algorithms decays rapidly with the dimension of  $S$ . This motivates finding low dimensional summary statistics. However, it is crucial that these are also informative, as otherwise the problem of inaccurate results described by Robert et al. (2011) can occur.

This paper sets out a method to choose  $S(x)$  for use in model selection. We give a theoretical result showing the existence of a low dimensional vector of statistics sufficient for model choice (under an appropriate definition given later). Our method aims to estimate such a vector. The idea is to use an extra simulation step to produce many  $(\mathcal{M}, \theta, x)$  triples and then fit simple regression models of  $\mathcal{M}$  on  $x$ . Predictors from the fitted regressions form estimates of low dimensional sufficient statistics, and are used as  $S$  in a main ABC analysis. We refer to the approach as the *semi-automatic method* as it adapts the method of the same name in Fearnhead and Prangle (2012) which chooses  $S$  by regressing  $\theta$  on  $x$  when the aim is inference of continuous parameters.

We expect that the targeted sufficient statistics are often complicated functions of the data which are hard to estimate globally. To make the task easier, we advise that the regressions are based on data simulated, within each model, from a limited subset of parameter values which is judged by preliminary analysis to hold most of that model’s posterior mass. In other words, the simulation step mentioned above performs simulations from the models of interest following a truncation of their parameter supports. The resulting  $S$  can only be expected to perform well for choice between these truncated models. A separate theoretical contribution of the paper is to relate results from such a choice to the original model choice problem.

Our approach of performing regressions based on simulated data is similar to Estoup et al. (2012) who instead use linear discriminant analysis. We expect our other contributions would also be useful to this approach. Other work on ABC summary statistics has focused on validating a particular choice of  $S$ . One approach is to run ABC analyses on a large number of simulated data sets to check whether  $S$  provides accurate results (Sousa et al., 2012; Sjödin et al., 2012). Marin et al. (2012) give a complementary approach, identifying necessary and sufficient properties of  $S$  for an ABC analysis to be consistent in an asymptotic regime corresponding to highly informative data. Essentially,  $S$  must have different asymptotic means under the models. Given a choice of  $S$ , this property can be tested theoretically or through simulation. Validation techniques are useful, but not sufficient, to choose  $S$  for high dimensional genetic data where it is infeasible to compare all possible choices of  $S$ . Our contribution is a method which can be applied in this setting to propose good choices of  $S$ .

Ideally the same ABC simulations would be used to provide inference on models and also their parameters. The method we present provides summary statistics suitable for model choice only. It would be desirable to augment them with informative summaries on model parameters, and we give an approach to do this that is specific to our main application. General methods are an interesting topic for future research.

The remainder of the paper is organised as follows. Section 2 describes ABC methods and our notation. Section 3 gives theoretical results on sufficiency, with proofs delayed until an appendix. Section 4 explains our semi-automatic ABC method, and Section 5 illustrates it for simple examples. The application to *Campylobacter* data is given in Section 6, and the article concludes with a discussion in Section 7. Further theoretical and simulation results are provided as supplementary material (Prangle et al., 2013).

## 2 Background

Denote by  $\mathcal{M}$  a random variable which can take values  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_M$ , representing possible models. Let  $p_M$  be its prior mass function. In an abuse of notation  $\mathcal{M}$  will also denote a generic value of the variable, with usage clear from the context. Each model represents a joint distribution  $\pi(x, \theta | \mathcal{M})$  on the data  $x$  and parameters  $\theta \in \Theta$ . This can be written as the product of prior and likelihood terms but we concentrate on the joint form for later convenience and to emphasise that the definition of a model includes a parameter prior. Note that it is possible for the parameters under each model to belong to different spaces, in which case  $\Theta$  is their union, and that  $\theta$  will also be used to denote both a random variable and generic value.

Bayesian inference concentrates on  $\pi(\theta | x, \mathcal{M})$  – the posterior distribution of parameters under a specific model – and  $\Pr(\mathcal{M} | x)$  – the posterior model probabilities. Inference on models can also be summarised using *Bayes factors*  $B_{ij} = \pi(x | \mathcal{M}_i) / \pi(x | \mathcal{M}_j)$ ; the ratio of the *evidences* under models  $\mathcal{M}_i$  and  $\mathcal{M}_j$ . The Bayes factor does not involve  $p_M$ , but incorporating this information allows calculation of the ratio of posterior weights:

$$\Pr(\mathcal{M}_i | x) / \Pr(\mathcal{M}_j | x) = B_{ij} p_M(\mathcal{M}_i) / p_M(\mathcal{M}_j).$$

ABC is used in situations where it is possible to simulate  $x | \mathcal{M}, \theta$  but evaluation of the density  $\pi(x | \mathcal{M}, \theta)$  is impossible or impractically costly. A simple approach to ABC inference is Algorithm 1 (Grelaud et al., 2009).

---

**Input:** Observed data  $x_{\text{obs}}$ , and a function  $S(\cdot)$ .  
A threshold  $h \geq 0$  and a distance function  $d(\cdot, \cdot)$ .  
An integer  $N > 0$ .

**Iterate:** For  $i = 1, \dots, N$

1. Simulate  $\mathcal{M}^*$  from  $p_M(\mathcal{M})$ .
2. Simulate  $\theta^*$  from  $\pi(\theta | \mathcal{M}^*)$ .
3. Simulate  $x_{\text{sim}}$  from  $\pi(x | \theta^*, \mathcal{M}^*)$ .
4. Accept  $(\mathcal{M}^*, \theta^*)$  if  $d(S(x_{\text{obs}}), S(x_{\text{sim}})) \leq h$ .

**Output:** A set of accepted model and parameter pairs of the form  $(\mathcal{M}^*, \theta^*)$ .

---

Algorithm 1: Rejection sampling ABC incorporating model choice and parameter inference.

Letting  $\mathbb{I}$  represent an indicator function, define

$$p_{\text{ABC}}(\mathcal{M}|S(x)) \propto p_M(\mathcal{M}) \int \pi(S(x)|\mathcal{M}) \mathbb{I}[d(S(x_{\text{obs}}), S(x)) \leq h] dx,$$

$$\pi_{\text{ABC}}(\theta|\mathcal{M}, S(x)) \propto \pi(\theta) \int \pi(S(x)|\theta, \mathcal{M}) \mathbb{I}[d(S(x_{\text{obs}}), S(x)) \leq h] dx.$$

Then the sample of  $(\mathcal{M}, \theta)$  values output by Algorithm 1 is drawn from a distribution with conditionals  $\pi_{\text{ABC}}(\theta|\mathcal{M}, S(x))$  and marginal  $p_{\text{ABC}}(\mathcal{M}|S(x))$ .

In the limit  $h \rightarrow 0$ , the ABC target distributions just defined converge on  $\Pr(\mathcal{M}|S(x))$  and  $\pi(\theta|\mathcal{M}, S(x))$ . However, reducing  $h$  decreases the output sample size, increasing Monte Carlo approximation error. A *curse of dimensionality* result reviewed in the supplementary material shows that the rate of increase in error rises with the dimension of  $S$ . This motivates a low dimensional  $S$ . It is also important that  $S$  is informative so that the limiting ABC targets approximate the posterior distributions  $\Pr(\mathcal{M}|x)$  and  $\pi(\theta|M, x)$  well. Hence  $S$  is a crucial tuning choice.

In practice, the results of Algorithm 1 can be highly variable if some prior model masses are small. Algorithm 2 is a more stable alternative suggested by Grelaud et al. (2009).

---

As Algorithm 1 except:

1. Set  $\mathcal{M}^*$  to  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_M$  with equal probability.
- 

Algorithm 2: A more stable modification of Algorithm 1.

Algorithm 2 samples  $\mathcal{M}^*$  values from a uniform distribution rather than  $p_M$ , and it is necessary to correct the results to take this into account. Let  $n_i$  be the number of occurrences of  $\mathcal{M}_i$  in the output sample. Then  $n_i/n_j$  is an estimator of the Bayes factor  $B_{ij}$  and  $n_i p_M(\mathcal{M}_i) / \sum_{j=1}^M n_j p_M(\mathcal{M}_j)$  is an estimator of  $\Pr(\mathcal{M}_i|S(x))$ . The asymptotic and curse of dimensionality results outlined above continue to hold. See Grelaud et al. (2009) and the supplementary material for full details.

More efficient ABC model choice algorithms have been proposed, mainly based on sequential Monte Carlo (SMC) (e.g. Toni and Stumpf, 2010; Del Moral et al., 2012). However, the tuning issues just described remain. The SMC algorithm of Toni and Stumpf (2010) is used later and described in the supplementary material. Another approach to improve the quality of ABC results is to *post-process* them. This uses accepted parameters  $\theta^{*,1}, \theta^{*,2}, \dots$ , models  $\mathcal{M}^{*,1}, \mathcal{M}^{*,2}, \dots$  and the corresponding simu-

lations  $x^{*,1}, x^{*,2}, \dots$ . For parameter inference *regression adjustment* (Beaumont et al., 2002; Blum and François, 2010) fits a model  $\theta = f(x, e)$ , where  $f$  is a deterministic function and  $e$  a random residual, and outputs adjusted values  $\theta'^i = \hat{f}(x_{\text{obs}}, \hat{e}^i)$ . Model choice results can be post-processed by fitting a multinomial regression model  $\Pr(\mathcal{M}|x) = g(x)$  and returning  $\hat{g}(x_{\text{obs}})$  (Beaumont, 2008).

### 3 Theory

A statistic  $S(x)$  of data  $x$  is said to be *Bayes sufficient* for parameter  $\theta$  if  $\theta|S(x)$  and  $\theta|x$  have the same distribution for any prior distribution and almost all  $x$  (Kolmogorov, 1942). This is a natural definition of sufficiency for ABC, as it shows that in an ideal ABC algorithm with  $h \rightarrow 0$ , the ABC target distribution equals the correct posterior when  $S$  is used. Throughout later sections of this paper we use “sufficient” to mean Bayes sufficient. Theorem 1 gives an alternative characterisation of Bayes sufficiency for  $\mathcal{M}$  in the setting described in Section 2.

**Theorem 1** *Let  $T(x) = \{T_1(x), T_2(x), \dots, T_{M-1}(x)\}$  where*

$$T_i(x) = \Pr(x|\mathcal{M}_i) / \left[ \sum_{j=1}^M \Pr(x|\mathcal{M}_j) \right].$$

*Then  $S$  is Bayes sufficient for  $\mathcal{M}$  if and only if there exists a function  $g$  such that  $g[S(x)] = T(x)$  for almost all  $x$ .*

Theorem 1 shows that for any situation with  $M$  models there are sufficient statistics for model choice of dimension  $M - 1$ , namely the vector  $T(x)$ . Furthermore, vectors  $S(x)$  which can be transformed to  $T(x)$  are also sufficient.

**Proof** See Appendix.

A sketch of the proof is as follows. The theorem states that Bayes sufficiency of  $S(x)$  for  $\mathcal{M}$  is equivalent to there being a deterministic transformation from  $S(x)$  to  $T(x)$ . The latter vector is  $M - 1$  posterior probabilities given observations  $x$  and uniform  $p_M$ . Under uniform  $p_M$ , conditioning  $\mathcal{M}$  on  $S(x)$  satisfying this condition clearly recovers the posterior weights. Reweighting can be used to show that the posterior is also recovered under any other  $p_M$ . The converse can be shown by construction.

One particular sufficient choice of  $S(x)$  used later is a vector of all Bayes factors under a one-to-one transformation. Additionally, we note that a sufficient  $S(x)$  may

contain summaries which do not contribute to  $T(x)$  but are useful for parameter inference.

Theorem 1 is similar to Theorem 3a of Fearnhead and Prangle (2012), which shows that for continuous parameters  $\theta$ ,  $S(x) = E(\theta|x)$  is an optimal choice to estimate parameter means in terms of minimising quadratic error loss. However this  $S(x)$  is typically not sufficient for  $\theta$ . Theorem 1 is a stronger result for the case of model choice (or, equivalently, for estimating discrete parameters) showing the existence of low dimensional vectors of sufficient statistics.

## 4 Method

The low dimensional sufficient statistics described by Theorem 1 are generally not available. However their existence motivates an approach of approximating them from simulated data, and then using these approximations as  $S(x)$  within ABC, as outlined in Algorithm 3. Step 2 requires some user input, as will be described in Section 4.1, so the method is referred to as “semi-automatic ABC”.

- 
1. Simulate a large number of  $(\mathcal{M}, \theta, x)$  triples.
  2. Calculate  $S(x)$  by estimating sufficient statistics from simulations.
  3. Perform the ABC analysis using  $S(x)$ .
- 

Algorithm 3: Outline of simple semi-automatic ABC for model choice. Full details of the steps are given in Sections 4.1 and 4.2.

Sufficient statistics are likely to be highly complicated functions of the data due to the complexity of the models, and thus hard to approximate. To make the task more tractable, we recommend some optional extra steps to give Algorithm 4. This simplifies the models by concentrating on the most likely parameter values. We view this as replacing the models  $\pi(\theta, x|\mathcal{M}_i)$  with *truncated models*

$$\pi(\theta, x|\mathcal{M}'_i) \propto \pi(\theta, x|\mathcal{M}_i)\mathbb{I}(\theta \in R_i), \quad (1)$$

where  $R_i$  is a *training region* for model  $\mathcal{M}_i$ . Calculation of  $S$  is performed using data simulated from the truncated models. The resulting  $S$  estimates sufficient statistics for the choice between the truncated rather than original models. Therefore the main ABC analysis must be performed between the truncated models, and, as will be shown

in Section 4.2, the results can be used to estimate the model choice posterior for the original problem.

- 
1. Perform an ABC *pilot analysis* with ad-hoc summary statistics. Use the output for each model to choose training regions  $R_i$  of parameters which contain most of the posterior probability for each model  $\mathcal{M}_i$ .
  2. Simulate a large number of  $(\mathcal{M}, \theta, x)$  triples using truncated models.
  3. Calculate  $S(x)$  by estimating sufficient statistics from simulations.
  4. Perform the ABC *main analysis* using  $S(x)$  and truncated models.
  5. Use truncation correction to estimate posterior probabilities.
- 

Algorithm 4: Semi-automatic ABC for model choice with truncation steps. Full details of the steps are given in Sections 4.1 and 4.2.

The remainder of this section discusses the implementation of the steps in these algorithms in more detail. Performance is assessed through simulation examples in Section 5.

## 4.1 Calculating summary statistics

This section describes a logistic regression based approach to estimating sufficient statistics from simulated *training data*. A motivating example is the case of two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , with training data drawn from the joint distribution on  $(\mathcal{M}, x)$ , where  $x = (x_1, x_2, \dots, x_p)$ . Define  $q(x) = \Pr(\mathcal{M}_1|x)$ . This is clearly a sufficient statistic for  $\mathcal{M}$ . Logistic regression can be used to fit

$$\text{logit } q(x) := \log\{q(x)/[1 - q(x)]\} = \beta_0 + \sum_{i=1}^p \beta_i x_i. \quad (2)$$

The fitted  $\hat{q}(x)$  is an estimate of a sufficient statistic. Note also that  $q(x)/[1 - q(x)]$  is the Bayes factor multiplied by a constant depending on the prior model weights.

To improve on the fit of (2) and cope with situations where  $x$  is very high dimensional or not a fixed-length vector, in practice we fit instead

$$\text{logit } q(x) = \beta^T f(x), \quad (3)$$

where  $f(x)$  is a vector of transformation of  $x$ , including a constant term. This can



perform initial dimension reduction and introduce non-linear functions of the data into the regression. Example choices of  $f(\cdot)$  used later are 1) order statistics of raw data 2) a large number of summaries of genetic sequence data used in previous literature, and transformation of these (a constant term is also included in both cases). To assist in the choice of  $f(\cdot)$ , regression diagnostics can be used, for example to compare the quality of the logistic regression fits for some  $f_1(\cdot)$  and  $f_2(\cdot)$ . The supplementary material gives examples in which cross-validation estimates of the deviance are compared.

In general the aim is to calculate  $S$  for choice between models  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_M$ , which for this discussion may represent original or truncated models. Fix a pair of distinct models,  $\mathcal{M}_i$  and  $\mathcal{M}_j$ , and consider the subset of training data made up of only the simulations from these models. Logistic regression can be used as above to estimate the probability  $q_{ij}$  of  $\mathcal{M}_i|x$  under the  $(\mathcal{M}, x)$  distribution for this training data subset. This is repeated for each pair of distinct models, and results in a vector of one-to-one transformations of Bayes factors. This target was shown to be sufficient for  $\mathcal{M}$  in Section 3.

The logistic regression method set out above gives  $\dim(S) = M(M-1)/2$ , whereas Theorem 1 shows there are sufficient statistics of dimension  $M-1$ . Alternative regression methods can be used to give  $\dim(S) = M-1$ , for example estimating an appropriate subset of the Bayes factors or multinomial regression. In this paper we consider only examples with  $M \leq 3$  so the logistic regression approach has limited excess dimension. We believe it also aids robustness. Even if the logistic regression for one pair of models fits poorly (as is the case in the *Campylobacter* application), the others can still allow a good overall estimate of sufficient statistics.

## 4.2 Other steps

**Pilot analysis** The pilot ABC analysis uses an ad-hoc choice of summary statistics  $S_{\text{pilot}}$ . The purpose of the pilot analysis is to roughly identify regions containing most of the posterior mass, so the procedure should be reasonably robust to the choice of  $S_{\text{pilot}}$ . Fearnhead and Prangle (2012) illustrate this argument by example. Validation tests could also be performed to test the quality of ABC output from analysing simulated data using  $S_{\text{pilot}}$ .

In our implementation the pilot uses an ABC model choice algorithm such as Algorithm 2. An alternative approach would be to perform a separate pilot run for each model, focusing only on finding training regions, rather than initial model choice analysis. We did not investigate this as a pilot analysis incorporating model choice

has useful properties. The estimated posterior can serve as a verification that the final results appear sensible. Also, if the pilot results are sufficiently convincing in showing that certain models are incompatible with the data, they could be ruled out at this stage saving computational resources.

**Training region choice** The training region  $R_i$  for model  $\mathcal{M}'_i$  should cover most of the posterior mass. Our implementation is to choose a hypercube, with the range of each parameter being the interval of sampled values.

**Simulating data** We generate training data from the distribution on  $(\mathcal{M}, \theta, x)$  defined by the priors and models (or truncated models). An alternative model distribution can be used without affecting the arguments in Section 4.1 showing that the fitted summary statistics are estimates of sufficient statistics. This would be useful if some prior model weights are too small to fit all regressions well.

**Truncation correction** Results of the main ABC analysis choosing between truncated models can be used to estimate those for the original model choice problem by the following consequence of (1):

$$\pi(x|\mathcal{M}_i) = r_i \pi(x|\mathcal{M}'_i), \quad \text{where } r_i = \Pr(\theta \in R_i|\mathcal{M}_i) / \Pr(\theta \in R_i|x, \mathcal{M}_i).$$

That is, the evidence of  $\mathcal{M}_i$  equals that of  $\mathcal{M}'_i$  multiplied by  $r_i$ , the ratio of the prior and posterior probabilities of  $R_i$ . This allows estimation of Bayes factors or posterior probabilities for the original models given  $r_i$  values. As  $R_i$  is chosen with the aim of containing most of the posterior mass, we estimate its posterior probability by 1, giving an estimate  $\hat{r}_i = \Pr(\theta \in R_i|\mathcal{M}_i)$ . This prior probability can usually be calculated directly when  $R_i$  is a hypercube.

## 5 Examples

To illustrate our semi-automatic ABC method, we apply it to three simple binary model selection examples from the literature (Didelot et al., 2011; Marin et al., 2012), and extend one of these to a 3 model example. The examples are summarised in Table 1. The binary examples are the first two models in each letter group, and the 3 model example is the full A group. In each case the data are 100 independent draws from one of the models and the models have equal prior probabilities. All ABC analyses were performed using Algorithm 2.

Name	Model	Prior
A1	Poisson( $\theta$ )	$\theta \sim \text{Exponential}(1)$
A2	Geometric( $\theta$ )	$\theta \sim \text{Uniform}(0, 1)$
A3	Binomial(10, $\theta$ )	$\theta \sim \text{Beta}(1, 9)$
B1	Laplace( $\theta, 1/\sqrt{2}$ )	$\theta \sim \text{Normal}(0, 2^2)$
B2	Normal( $\theta, 1$ )	$\theta \sim \text{Normal}(0, 2^2)$
C1	gk(0, 1, 0, $k$ )	$k \sim \text{Unif}(-0.5, 5)$
C2	gk(0, 1, $g, k$ )	$(g, k) \sim \text{Unif}([0, 4] \times [-0.5, 5])$

Table 1: Models used in the examples of Section 5. For details of the  $g$ -and- $k$  distribution see Rayner and MacGillivray (2002).

**Binary model selection** The semi-automatic ABC method of Algorithm 4 was implemented starting with a pilot analysis using  $S_{10}(x) = (x^{(5)}, x^{(15)}, \dots, x^{(95)})$  where  $x^{(i)}$  is the  $i$ th order statistic. Model choice summary statistics were fitted as described in Section 4.1 using  $f(x) = (1, x^{(1)}, x^{(2)}, \dots, x^{(100)})$ . No other summaries were added for parameter inference. The analysis used  $2 \times 10^4$  simulations, one quarter for the pilot and the rest used for both summary statistic fitting and the main analysis. The pilot and main analysis both accepted 100 simulations. Some alternative ABC analyses on the data were performed, each using the same total number of simulations and acceptances. Firstly, the analysis was repeated using Algorithm 3. Secondly, standard ABC analyses were performed with Algorithm 2 using (a)  $S = S_{10}$  (b)  $S$  as in Marin et al. (2012); 4th and 6th moments for B, 10% and 90% quantiles for C. All ABC analyses used the following distance

$$d(x, y) = \left[ \sum_{i=1}^p (x_i - y_i)^2 / \hat{\sigma}_i^2 \right]^{1/2}, \quad (4)$$

i.e. Euclidean distance between  $p$ -dimensional summary statistics normalised by estimated standard deviations,  $\hat{\sigma}_i$ . The latter were estimated from the simulated data.

Figure 1 shows estimated posterior probabilities for  $S_{10}$  and Algorithm 4. Numerical summaries of estimation quality are given in Table 2. This reports the entropic loss (Robert, 1996),

$$-\sum_{i=1}^{100} \log \hat{\text{Pr}}(m_{0,i} | x_{\text{obs},i}),$$

the negative log probability of the correct models  $m_{0,1}, \dots, m_{0,100}$  estimated from the corresponding simulated datasets  $x_{\text{obs},1}, \dots, x_{\text{obs},100}$ . Also reported is the misallocation rate; the proportion of datasets where the highest weighted model was not the correct

model. Our method provides an improvement in all scenarios, although this is modest for example C. The use of the truncation steps from Algorithm 4 is shown to sometimes be crucial; when Algorithm 3, which omits these, is used instead, the results for example C are the worst of all methods. However the effect is problem dependent; in example B it made little difference. Exact posterior calculations are possible for examples A and B (the required Laplace marginal likelihood calculations are described in Appendix 1 from version 1 of Marin et al. 2012), and in both cases Algorithm 4 provides comparable results.

We attempted to apply post-processing by the method of Beaumont (2008). For example A this was usually not possible as there was no variation in the accepted summaries, which were discrete in this case, or because all acceptances were for a single model. For the other examples, it had little effect on entropic loss or misallocation rate, so these are not reported.

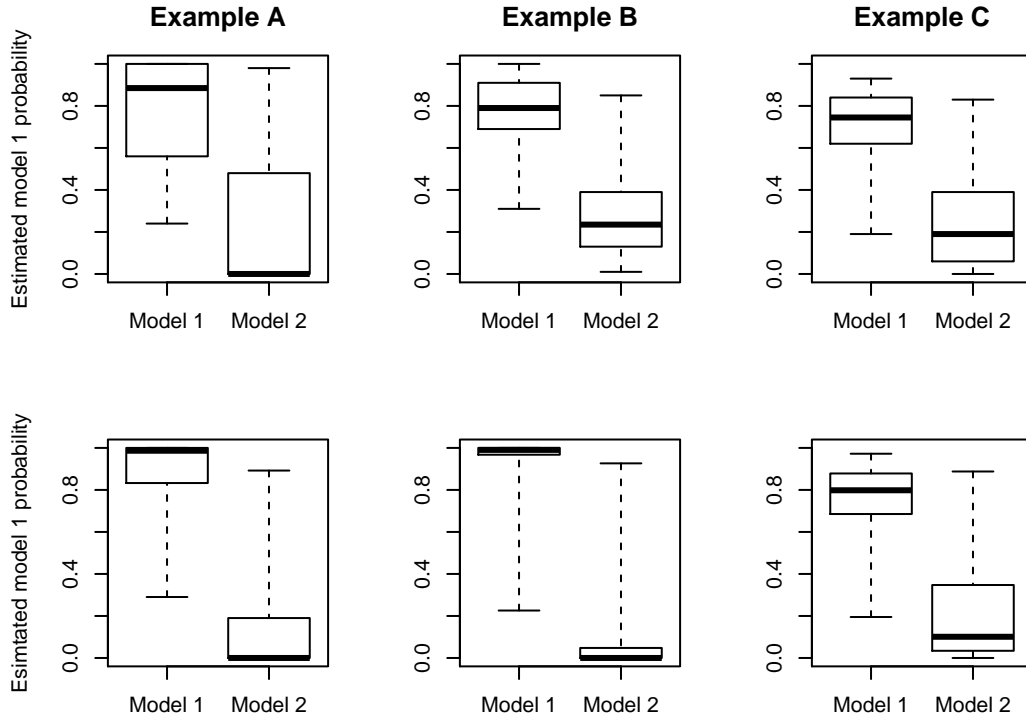


Figure 1: Boxplots of posterior probabilities of model 1 estimated by ABC (without post-processing) for 100 simulated datasets in each of three binary model comparison examples. The boxplots show quartiles of the values. Within each graph results are split by which model generated the data. The top row uses  $S = S_{10}$ , and the second row chooses  $S$  by semi-automatic ABC (Algorithm 4). The columns represent three model choice examples detailed in Table 1.

Summary statistics	Example			
	Binary A	Binary B	Binary C	3 models
$S_{10}$	33.0 (17%)	33.5 (11%)	43.0 (16%)	70.7 (39%)
From literature	-	55.3 (25%)	40.9 (20%)	-
From Algorithm 3	30.2 (14%)	13.5 (5%)	$\infty$ (21%)	65.9 (42%)
From Algorithm 4	19.8 (15%)	13.9 (7%)	38.4 (14%)	58.9 (33%)
Posterior	19.8 (12%)	15.6 (8%)	-	58.1 (36%)

Table 2: Entropic loss and misallocation rate (in brackets) from several ABC analyses of 100 simulated datasets in each of four model comparison examples, detailed in Table 1. The final row shows values under the exact posterior, where these are available, for comparison.

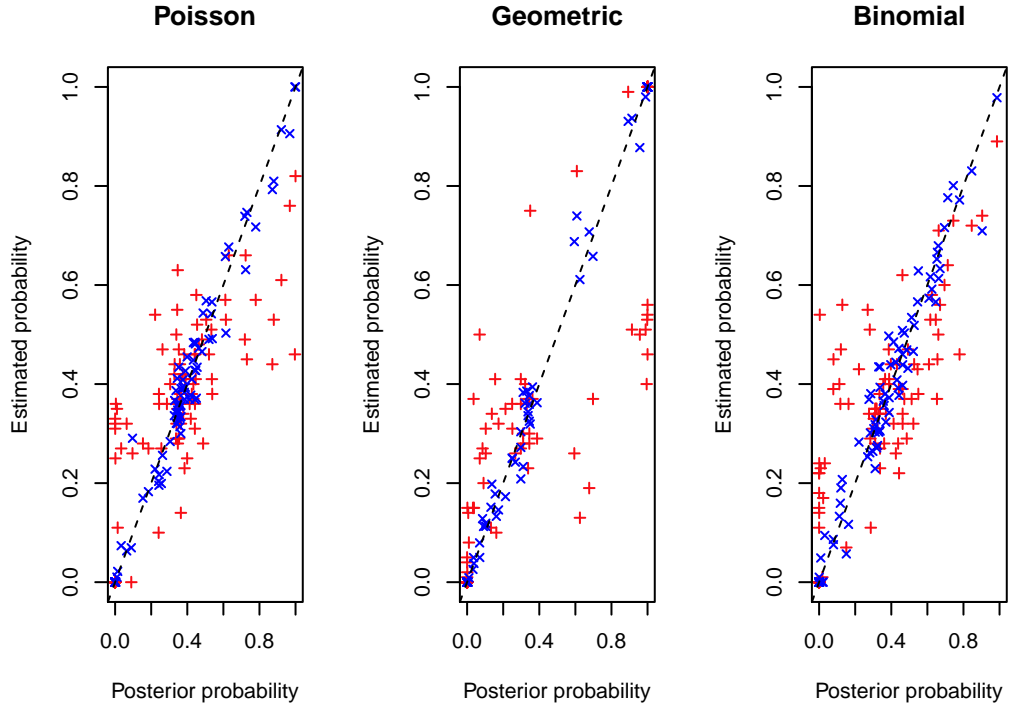


Figure 2: Plots of true posterior model weight against ABC estimates for 100 simulated datasets in a three model example. Pluses are for  $S = S_{10}$  and crosses for  $S$  chosen by semi-automatic ABC (Algorithm 4).

**Selection between three models** Algorithms 3 and 4 were implemented as for the two model examples, with the addition that three summary statistics were fitted, corresponding to three pairs of models. Figure 2 plots exact posterior probabilities against ABC estimates, and shows that Algorithm 4 performs better than the comparison analysis using  $S = S_{10}$ . This is confirmed by the quantitative summaries in Table 2, which also shows that Algorithm 4 outperforms Algorithm 3 and achieves comparable results to the true posterior values. Post-processing results are not shown because, as mentioned above, they could usually not be calculated for this example.

## 6 Application

*Campylobacter jejuni* and *C. coli* are bacterial pathogens that are a major cause of human gastroenteritis around the world (Humphrey et al., 2007). They are considered commensals of a wide variety of animals, including poultry, ruminants and wild birds, and human infection occurs as a result of ingesting contaminated food or drinking water and via direct contact with animal faeces (Savill et al., 2003). New Zealand has very high rates of campylobacteriosis and an investigation into the source of human infection (Mullner et al., 2009) has generated a large dataset of isolates from humans and animals that have been characterized by multilocus sequence typing, MLST (Dingle et al., 2001). The dataset of *C. jejuni* and *C. coli* isolates from New Zealand has been used to inform control policy (Sears et al., 2011) and to estimate evolutionary parameters, such as the rates of mutation and recombination (Yu et al., 2012). We focus on the question of demographic history, which is of particular interest in New Zealand due to the relatively recent colonization by man and the unique pattern of animal introductions (both wildlife and domestic animals) (Atkinson and Cameron, 1993). We ask: can we detect historic growth in the effective population size, and if so, does it correspond to a particular historical event? The relative isolation of this location means that neglecting ongoing exchange with outside populations is reasonably realistic. MLST data are available for over 3000 isolates from a variety of hosts.

We present our methods and results below, with a discussion given in Section 7.2. Some further details are provided as supplementary material.

## 6.1 Models and priors

We modelled the *C. jejuni* data using a coalescent model using the Jukes-Cantor model of DNA substitution and incorporating the gene conversion recombination model of Wiuf and Hein (2000) with exponential demographic growth, as simulation of this scenario is straightforward using existing tools (detailed below). However, simulation of a large dataset is prohibitively slow so we used a random subsample of 100 isolates. Coalescent theory suggests that such a sample size captures much of the information of the full sample (Nordborg, 2004), and simulation based checks on informativeness are detailed in the supplementary material. The selected isolates were confirmed to be *C. jejuni* using the PubMLST database and through a phylogeny analysis of these isolates and a representative *C. coli* sequence. Three models were considered, with equal prior weights: *Model 1* no growth; *Model 2* growth for 50 years (since expansion of the New Zealand poultry industry); *Model 3* growth for 170 years (since introduction of European livestock, primarily from Australia and the UK).

Each model has three biological parameters: a recombination rate, mean track length (i.e. length of recombining DNA segment) and mutation rate. Models 2 and 3 also have a growth parameter. To aid interpretability we parameterised this as the relative increase in the effective population size during the period of growth. Prior information on parameters is summarised in Table 3. Mutation and recombination rates are given per kilobase per  $2N_e g$  years, where  $N_e$  is the effective population size and  $g$  the generation length in years. Wilson et al. (2009) estimated the mean time to coalescence,  $N_e g$ , at 218 with an interval estimate of [155, 288]. To simplify our model, we fix  $N_e g = 218$ . We expect that variations of  $N_e g$  within the quoted interval will not affect the detection of growth. Mean recombination length is in kilobase units. The relative growth parameter is unitless as it is a ratio of effective population sizes.

Growth priors are based on demographics of the principal host; poultry for model 2 and sheep/cattle for model 3. Rough estimates of host growth rates are used, based on the data of Binney et al. (2013), with variance increased to account for uncertainty of the link between bacterial and host demographics. Biological parameter priors are based on analysis of other *C. jejuni* data in Wilson et al. (2009). This assumed a no growth model, so these priors may not be appropriate for models 2 and 3. Sensitivity analysis detailed in the supplementary material also considers a much less informative biological prior.

Parameter	Units	Model	Point estimate	95% CI	Log normal	
					Mean	Sd
Mutation rate	$kb^{-1}(2N_{eg})^{-1}$	All	13.7	[8.1, 23.2]	2.62	0.27
Recombination rate	$kb^{-1}(2N_{eg})^{-1}$	All	1.31	[0.03, 51.5]	0.27	1.87
Mean track length	$kb$	All	4.52	[0.1, 209.9]	1.51	1.96
Relative growth	-	2	4.06	[1.5, 10.8]	1.40	0.50
Relative growth	-	3	33.1	[2.9, 383.8]	3.50	1.25

Table 3: Details of the parameter priors used in Section 6. Priors are assumed to be the product of a log normal prior for each individual parameter. The point estimates are geometric means. The recombination length prior was truncated below 1 base pair, and the recombination rate above  $25kb^{-1}(2N_{eg})^{-1}$  to avoid excessively slow simulations (All estimated posteriors for recombination rate were well below this - see Figure 2 of supplementary material.)

## 6.2 Methods

Data sets were simulated using *ms* (Hudson, 2002) and *seq-gen* (Rambaut and Grassly, 1997). Genetic summaries required were calculated using R (R Core Team, 2012), which was also used to code the inference algorithms.

We implemented semi-automatic ABC (Algorithm 4) as follows. First a pilot analysis was performed using the ABC SMC algorithm of Toni and Stumpf (2010) (detailed in the supplementary material) with 1000 particles. This targeted log-transformed parameters, as on the original scale the target distribution is roughly log-normal and hard for the algorithm to explore. The summary statistics were a set of 15 genetic summaries (these, and other summaries used below, are listed in the supplementary material). The distance function was Equation (4), Euclidean distance between normalised summary statistics, with standard deviations estimated from 100 datasets simulated from the prior predictive distribution. These simulations were also used to choose an initial ABC threshold: the median of the distances between these datasets and the observations. In following SMC iterations, the threshold was the median of distances for accepted particles in the preceding step. The algorithm terminated after the iteration which reached  $2 \times 10^4$  simulated data sets.

To fit summary statistics,  $2 \times 10^4$  datasets were simulated using the training regions. Model choice summaries were fitted as described in Section 4.1 and summaries for parameter inference by linear regression (detailed shortly). For all regressions the vector of covariates  $f(\cdot)$  consisted of 3 cubic B-spline bases for each of 125 genetic summaries, giving a total of 375 covariates, and a constant term. Spline transformations were included to capture non-linear effects. Due to the large number of covariates,  $L_1$  penalised versions of logistic and linear regression were used, using the ‘glmnet’ R package (Friedman et al., 2010) with the tuning parameter chosen by



cross-validation. Cross-validation estimates of fitting error were used to investigate which genetic summaries were most informative and to validate many of our modelling and tuning choices (details in supplementary material).

Exploratory analysis showed that for each parameter a single estimator could perform reasonably well under all models (details in supplementary material). To keep  $\dim(S)$  small, our  $S$  is the concatenation of such estimators with model choice statistics. A single hypercube training region was used for all models to prevent behaviour of a particular model being overrepresented in any region of parameter space. This training region was the product of the parameter ranges from the entire pilot output, regardless of model. The regression responses were log-transformed parameters, supported by exploratory analysis of Box-Cox transformations. The resulting predictors were exponentiated to use in  $S$ . Regressions for biological parameters were fitted using the simulations from all models, while those for the demographic parameter used simulations from the growth models only.

The final  $S$  vector used in the main ABC analysis consisted of four parameter estimators and three statistics for model choice. The analysis used the distance function (4) with summary statistic standard deviations estimated from the training data. The analysis used the same SMC ABC algorithm as the pilot run, again with 1000 particles and targeting log-transformed parameters. The initial threshold was the median of distances to the observed data calculated from the training data, with subsequent thresholds chosen as in the pilot run. The algorithm terminated after the iteration which reached  $4 \times 10^4$  simulated data sets.

### 6.3 Results

Table 4 summarises the model choice results for the pilot and main analyses, including the effect of regression post-processing as in Beaumont (2008). They agree in putting the majority of the weight on model 1, the no growth model. Effective sample sizes (Liu, 1996) show that Monte Carlo error is approximately equal to that of a moderately large independent sample. The supplementary material details sensitivity analyses which vary the parameter priors and the subsample of isolates used as observations. With the exception of some pilot analyses, the weight placed on model 1 remains in the range 80 – 100%. ABC analyses of simulated datasets are also described in the supplementary material. Although only a small number were possible due to the high computational cost, the results suggest that the analyses are capable of distinguishing the no-growth from the growth models, with the main analysis doing so more accurately.

Analysis	ESS	Post-processed?	Model 1	Model 2	Model 3
Pilot	348	No	0.86	0.11	0.04
		Yes	1.00	0.00	0.00
Main	600	No	0.96	0.03	0.01
		Yes	0.92	0.03	0.05

Table 4: Estimated posterior probabilities and effective sample sizes from ABC analyses on *Campylobacter* data.

Table 5 summarises the parameter inference results. Marginal density plots are provided in the supplementary material. The table includes results from applying the regression adjustment of Beaumont et al. (2002) to model 1 output. This was not applied to other models as there were too few accepted particles to expect it to be stable. The most notable finding is the low estimate of recombination rate, discussed further in Section 7.2. Additionally, informative estimates are made for mutation rate and relative growth. The latter concentrates on low values, providing further evidence against significant growth. Sensitivity analyses detailed in the supplementary material support these qualitative findings, although the numerical values are less robust than those for model choice.

		Recombination rate $kb^{-1}(2N_{eg})^{-1}$	Mean track length $kb$	Mutation rate $kb^{-1}(2N_{eg})^{-1}$	Relative growth
Prior	Model 1	1.31 [0.03, 51.5]	4.52 [0.1, 209.9]	13.7 [8.1, 23.2]	4.06 [1.5, 10.8] 33.1 [2.9, 383.8]
	Model 2	1.31 [0.03, 51.5]	4.52 [0.1, 209.9]	13.7 [8.1, 23.2]	
	Model 3	1.31 [0.03, 51.5]	4.52 [0.1, 209.9]	13.7 [8.1, 23.2]	
Pilot	Model 1	0.34 [0.02, 5.21]	2.43 [0.06, 88.2]	11.4 [7.63, 16.7]	2.12 [1.07, 3.07] 4.81 [0.97, 19.0]
	Model 1 (adjusted)	0.18 [0.02, 1.87]	1.04 [0.05, 24.8]	12.8 [10.2, 16.7]	
	Model 2	0.28 [0.02, 2.45]	1.99 [0.09, 24.1]	12.6 [8.76, 17.2]	
	Model 3	0.17 [0.01, 0.78]	1.58 [0.08, 94.9]	12.2 [8.80, 15.1]	
Main	Model 1	0.55 [0.02, 3.74]	5.81 [0.17, 239.2]	12.9 [10.1, 16.5]	1.51 [0.85, 2.71] 1.12 [0.41, 2.44]
	Model 1 (adjusted)	0.22 [0.02, 1.18]	2.98 [0.22, 63.2]	13.0 [10.6, 15.9]	
	Model 2	0.24 [0.01, 3.53]	5.73 [0.52, 239]	14.0 [11.6, 16.5]	
	Model 3	0.34 [0.01, 3.37]	3.08 [0.40, 128]	12.6 [9.81, 16.4]	

Table 5: Parameter point estimates (geometric means) and 95% credible intervals from prior and ABC analyses on *Campylobacter* data.

The regression and ABC results were also used to find which genetic summaries were particularly informative, and to show that some aspects of the data fitted poorly under any model. These results are given in the supplementary material, and can inform future modelling and analyses.

## 7 Discussion

### 7.1 ABC Methodology

It is often desirable to perform model choice and parameter inference using the same simulations. Our methodology focuses on producing  $S$  appropriate for model choice only. Section 6 contains an application-specific example of adding a small number of further summaries to  $S$  which are informative for parameter inference. General purpose methods to choose such low dimensional summaries would be useful. However, often each model may require separate summaries, so that a choice of  $S$  suitable for model choice and parameter inference would be high dimensional. An alternative strategy is to develop ABC methods in which comparisons of simulated and observed data do not always use the same summaries. A simple approach would be to perform separate rejection sampling analyses for model choice and for parameter inference under each model. A possible alternative is an MCMC algorithm which moves between models using only summaries relevant to the model(s) involved in the current step.

There are numerous alternatives to logistic regression to fit summary statistics for model choice, such as linear discriminant analysis (Estoup et al., 2012) and a comparison of their performance within ABC may be interesting. Other parts of our semi-automatic method could also be varied. For example, our choice of  $S$  is a vector of one-to-one transformations of Bayes factors, and other transformations may perform differently. Also, other methods could produce a more accurate training region, such as fitting a flexible model to the pilot output.

For simplicity we have used relatively simple ABC algorithms. However, much progress is being made in improving algorithmic efficiency, especially of ABC SMC (e.g. Del Moral et al., 2012). Our work is complementary to this and it could be used with many such improved algorithms. Indeed ABC SMC algorithms can also be modified to incorporate semi-automatic ABC. For example, recall that in Section 5 the training data were reused as the simulations needed for ABC rejection sampling. As suggested by Barnes et al. (2012), in ABC SMC they could be similarly reused for the first SMC iteration.

### 7.2 *Campylobacter* application

Our main finding is support for a model with no change in the effective population size of *C. jejuni*. This is surprising over a period where its ecological niche has greatly increased. Analysis in the supplementary material shows some features of the data are

poorly fitted under all models, suggesting that more detailed demographic structure is necessary to fit the data well. One potential modification is subpopulation structure amongst the hosts which might reveal that only some support growing *C. jejuni* populations.

Our analysis also produced parameter estimates. Those for mutation rate and mean length of recombination tract are comparable to those from other work. The point estimates of recombination rate are somewhat smaller than those of Wilson et al. (2009), who performed a similar ABC analysis on a different dataset. Furthermore our credible intervals are much narrower, and exclude the estimates of Fearnhead et al. (2005), Biggs et al. (2011) and Yu et al. (2012), who find recombination and mutation rates to be of the same order of magnitude. The discrepancy with Wilson et al. (2009) is conceivably due to their use of a heavy tailed prior or ABC tuning differences such as choice of threshold. The others suggest differences in the model or data used. For example, as discussed by Yu et al. (2012), their analysis, and that of Biggs et al. (2011), is for closely related sequences, and may reveal a high level of recombination that is then removed by purifying selection.

**Acknowledgements** The authors acknowledge the Marsden Fund project 08-MAU-099 (Cows, starlings and *Campylobacter* in New Zealand: unifying phylogeny, genealogy, and epidemiology to gain insight into pathogen evolution) for funding this project. This publication made use of the *Campylobacter* Multi Locus Sequence Typing website (<http://pubmlst.org/campylobacter/>) developed by Keith Jolley and sited at the University of Oxford (Jolley and Maiden 2010, BMC Bioinformatics, 11:595). The development of this site has been funded by the Wellcome Trust.

## Appendix: Proof of Theorem 1

Bayes sufficiency of  $S(x)$  for  $\mathcal{M}$  is equivalent to the following being true for all  $i$  and  $p_M$ , and almost any  $x$ ,

$$\Pr(\mathcal{M}_i|S(x)) = \Pr(\mathcal{M}_i|x). \quad (5)$$

For convenience we introduce  $\mathbf{p} = (p_M(\mathcal{M}_i))_{1 \leq i \leq M}$  to represent the prior mass function. Also, let  $\mathbf{1}$  be a vector of  $M$  components equal to 1.

First assume  $S$  is Bayes sufficient for  $\mathcal{M}$ . Define  $h_i(S(x), \mathbf{p}) = \Pr_{\mathbf{p}}(\mathcal{M}_i|S(x))$  (i.e. the conditional probability under prior  $\mathbf{p}$ ) and note  $h_i(S(x), \mathbf{p}) = \Pr_{\mathbf{p}}(\mathcal{M}_i|x)$ . The required function is  $g(S(x)) = (h_i(S(x), M^{-1}\mathbf{1}))_{1 \leq i \leq M-1}$ .

It remains to prove Bayes sufficiency for  $\mathcal{M}$  given a function  $g$  of the form described in the theorem. Henceforth we consider only the case  $\mathbf{p} = M^{-1}\mathbf{1}$ , since in this case (5) is equivalent to  $\Pr(x|\mathcal{M}_i) = k \Pr(S(x)|\mathcal{M}_i)$  for some constant  $k$ , and applying Bayes' theorem to this proves (5) for general  $\mathbf{p}$ . It also suffices to show that (5) holds for all  $i < M$ ; the case  $i = M$  follows as probabilities sum to 1. Fix some  $i < M$  and define an indicator variable  $Y = \mathbb{I}[\mathcal{M} = \mathcal{M}_i]$ . Then  $T_i(x) = \Pr(\mathcal{M}_i|x) = E[Y|x]$  and  $\Pr(\mathcal{M}_i|S(x)) = E[Y|S(x)]$ . To prove (5), we will show that these conditional expectations are almost always equal. Standard properties of conditional expectation give  $E[Y|S(x)] = E[E\{Y|x\}|S(x)] = E[T_i(x)|S(x)]$ . Finally,  $E[T_i(x)|S(x)] = E[g_i(S(x))|S(x)] = g_i(S(x)) = T_i(x) = E[Y|x]$  for almost all  $x$  as required, where  $g_i(\cdot)$  represents the  $i$ th component of the  $g(\cdot)$  function.

## References

- Atkinson, I. A. and Cameron, E. K. (1993). Human influence on the terrestrial biota and biotic communities of New Zealand. *Trends in Ecology & Evolution*, 8:447–451.
- Barnes, C. P., Filippi, S., and Stumpf, M. P. H. (2012). Contribution to the discussion of Fearnhead and Prangle (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B*, 74:453.
- Beaumont, M. A. (2008). Joint determination of topology, divergence time, and immigration in population trees. In Renfrew, C., Matsumura, S., and Forster, P., editors, *Simulation, Genetics and Human Prehistory*, pages 134–154. McDonald Institute Monographs.
- Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035.
- Biggs, P. J., Fearnhead, P., Hotter, G., Mohan, V., Collins-Emerson, J., Kwan, E., Besser, T. E., Cookson, A., Carter, P. E., and French, N. P. (2011). Whole-genome comparison of two *Campylobacter jejuni* isolates of the same sequence type reveals multiple loci of different ancestral lineage. *PloS One*, 6(11):e27121.
- Binney, B., Biggs, P. J., Carter, P., Holland, B., and French, N. P. (2013). Historical livestock importation into New Zealand. *New Zealand Veterinary Journal*. (submitted).

- Blum, M. G. B. (2010). Approximate Bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187.
- Blum, M. G. B. and François, O. (2010). Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, 20:63–73.
- Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020.
- Didelot, X., Everitt, R. G., Johansen, A. M., and Lawson, D. J. (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis*, 6(1):49–76.
- Dingle, K. E., Colles, F. M., Wareing, D. R. A., Maiden, M. C. J., Ure, M. C. J., Maiden, R., Fox, A. J., Bolton, F. E., Bootsma, H. J., Willems, R. J., Urwin, R., and Maiden, M. C. (2001). Multilocus sequence typing system for *Campylobacter jejuni*. *Journal of Clinical Microbiology*, 39:14–23.
- Estoup, A., Lombaert, E., Marin, J.-M., Guillemaud, T., Pudlo, P., Robert, C. P., and Cornuet, J. (2012). Estimation of demo-genetic model probabilities with approximate Bayesian computation using linear discriminant analysis on summary statistics. *Molecular Ecology Resources*, 12(5):846–855.
- Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic ABC. *Journal of the Royal Statistical Society, Series B*, 74:419–474.
- Fearnhead, P., Smith, N. G. C., Barrigas, M., Fox, A., and French, N. (2005). Analysis of recombination in *Campylobacter jejuni* from MLST population data. *J Mol Evol*, 61:333–340.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).
- Grelaud, A., Robert, C., Marin, J.-M., Rodolphe, F., and Taly, J. F. (2009). ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, 4(2):317–336.
- Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338.

- Humphrey, T., O'Brien, S., and Madsen, M. (2007). Campylobacters as zoonotic pathogens: a food production perspective. *Int J Food Microbiol.*, 117(3):237–57.
- Kolmogorov, A. N. (1942). Determination of centre of dispersion and measure of accuracy from a finite number of observations (in Russian). *Izv. Akad. Nauk, USSR Ser. Mat.*, 6:332.
- Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6:113–119.
- Marin, J.-M., Pillai, N., Robert, C. P., and Rousseau, J. (2012). Relevant statistics for Bayesian model choice. *Preprint*. Available at <http://www.arxiv.org/abs/1110.4700>.
- Mullner, P., Spencer, S. E. F., Wilson, D. J., Jones, G., Noble, A. D., Midwinter, A. C., Collins-Emerson, J. M., Carter, P., Hathaway, S., and French, N. P. (2009). Assigning the source of human campylobacteriosis in New Zealand: A comparative genetic and epidemiological approach. *Infection, Genetics and Evolution*, 9(6):1311–1319.
- Nordborg, M. (2004). Coalescent theory. In *Handbook of statistical genetics*, volume 2. Wiley.
- Prangle, D., Fearnhead, P., Cox, M. P., Biggs, P. J., and French, N. P. (2013). Supplementary material for Semi-automatic selection of summary statistics for ABC model choice. Available at <http://www.arxiv.org/abs/???>
- R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rambaut, A. and Grassly, N. C. (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13:235–238.
- Rayner, G. D. and MacGillivray, H. L. (2002). Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, 12(1):57–75.
- Robert, C. P. (1996). Intrinsic losses. *Theory and decision*, 40(2):191–214.

- Robert, C. P., Cornuet, J. M., Marin, J.-M., and Pillai, N. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117.
- Savill, M., Hudson, A., Devane, M., Garrett, N., Gilpin, B., and Ball, A. (2003). Elucidation of potential transmission routes of *Campylobacter* in New Zealand. *Water Science and Technology*, 47(3):31–38.
- Sears, A., Baker, M. G., Wilson, N., Marshall, J., Muellner, P., Campbell, D. M., Lake, R. J., and French, N. P. (2011). Marked campylobacteriosis decline after interventions aimed at poultry, New Zealand. *Emerging infectious diseases*, 17(6):1007–1015.
- Sjödén, P., Sjöstrand, A. E., Jakobsson, M., and Blum, M. G. B. (2012). Resequencing data provide no evidence for a human bottleneck in africa during the penultimate glacial period. *Molecular Biology and Evolution*, 29(7):1851–1860.
- Sousa, V. C., Beaumont, M. A., Fernandes, P., Coelho, M. M., and Chikhi, L. (2012). Population divergence with or without admixture: selecting models using an ABC approach. *Heredity*, 108:521–530.
- Toni, T. and Stumpf, M. P. H. (2010). Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110.
- Wilson, D. J., Gabriel, E., Leatherbarrow, A. J. H., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Hart, C. A., Diggle, P. J., and Fearnhead, P. (2009). Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Molecular biology and evolution*, 26(2):385–397.
- Wiuf, C. and Hein, J. (2000). The coalescent with gene conversion. *Genetics*, 155:451–462.
- Yu, S., Fearnhead, P., Holland, B. R., Biggs, P., Maiden, M., and French, N. P. (2012). Estimating the relative roles of recombination and point mutation in the generation of single locus variants in *Campylobacter jejuni* and *Campylobacter coli*. *Journal of Molecular Evolution*, 74(5-6):273–280.